УДК 003’26

Г.К. ИСМАГИЛОВА,
кандидат филологических наук, преподаватель
Казанский федеральный университет
Р.В. БАГАУТДИНОВ,
студент
Казанский федеральный университет

МОДЕЛЬ НЕЙРОННОЙ СЕТИ ДЛЯ СБОРА ДАННЫХ ОБ ОШИБКАХ ПРИ ПЕРЕВОДЕ РАЗНОСТРУКТУРНЫХ ЯЗЫКОВ

Аннотация. Данная статья посвящена теоретической модели нейронной сети, при помощи которой можно отслеживать смысловые ошибки и неточности при переводе текстов между разноструктурными языками, для дальнейшей обработки.
Ключевые слова: Нейронные сети, теория Хомского, машинный перевод, сбор данных.
Abstract. This article is devoted to the theoretical model of neural network, through which is possible to track semantic errors and inaccuracies when translating texts of different lingual languages, for further processing.
Keywords: Neural networks, Chomsky theory, machine translation, data collection.

В данном докладе, мы предлагаем модель обработки данных, дополняющую существующие алгоритмы перевода текстов, призванную выявлять проблемы сопоставления структур языков, а также максимально устранить искажение смысловой характеристики текста при переводе, т.е. повысить адекватность перевода, не теряя при этом точности.

Наиболее популярной моделью устройства нейронной сети на данный момент, использующейся при машинном переводе, является SEQ2SEQ архитектура. Данная модель состоит из двух рекуррентных сетей: кодировщика и декодировщика. Задача кодировщика – получить из входной последовательности (изначального текста) представление по каждому слову, которое затем подаётся на декодировщик, который выполняет обратную задачу по генерации ответа на требуемом языке.

Однако, несмотря на сложную архитектуру, результат данной операции не всегда соответствует ожиданиям, и исходный смысл искажается тем сильнее, чем сильнее различаются структуры сопоставляемых языков.

Для решения этой проблемы, мы предлагаем задействовать дополнительную нейронную сеть для каждого языка, формирующую универсальную смысловую единицу, которая впоследствии будет использована для более глубокого сопоставления семантики исходных речевых единиц, и результата перевода.

Смысловой единицей в данном понимании является кортеж утверждений, сформированный в результате обработки текста. Данная конструкция формируется, опираясь на структурные особенности конкретного языка, учитывает лексику атомарных единиц языка и грамматические особенности, для формирования дополнительной информации, с целью последующего сопоставления [1, с.1].

Очевидно, что при формировании наборов утверждений необходимо учитывать лексические и грамматические особенности конкретного языка, поэтому для работы с каждым из требуемых языков необходима специально настроенная и обученная распознаванию ключевых моментов каждой речевой единицы сеть. В качестве топологии данной сети необходимо использовать feedback модель, по причине неоднозначности структуры построения речевых единиц, а в качестве уровней выделить типы данных единиц, и организовывать связи, ссылаясь на особенности структуры языка [3, с.21].

Данная идея была вдохновлена теорией Хомского об универсальной грамматике и формальным подходом языкознания. Суть идеи в том, чтобы, двигаясь от речевых конструкций к вложенному в них смыслу, достаточно упростив его, для возможности оперирования разноструктурными языками без искажения результата перевода, получить в итоге модель формирования смысловой структуры, вложенную в речевую единицу конкретного языка. Итоговое сопоставление данных моделей у двух речевых единиц различных языков поможет выявить, несут ли они схожую информацию, или нет.

Как итог, схема работы будет выглядеть следующим образом. В первую очередь, происходит стандартный перевод на целевой нейронной сети, которой требуется сбор данных о возможных ошибках сопоставления. Затем, каждый из исходного текста перевода и результата сопоставления, поступают на вход соответствующим их языкам нейросетям, формирующим смысловые единицы. И, наконец, происходит со-

поставление сформированных единиц, и запись результата сопоставления. Повторение подобной операции на различных вариациях входа позволит собрать информацию о наиболее часто возникающих ошибках, внести коррекции в существующие структуры нейронных сетей перевода (подразумевается, коррекция веса на связях нейронов вручную, но вполне возможна и более сложная модель интеграции дополнительной сети, в качестве автокорректирующей).

Конечно, формирование данной модели требует глубокого исследования целевых структур языков, и самих языков в частности, а сами смысловые единицы неспособны однозначно восстановить структуру исходного текста. Однако использование этой надстройки поможет отлавливать противоречия в смысловых конструкциях текста ДО, и ПОСЛЕ перевода. Так же она позволит собирать статистические данные о несоответствиях, и, в связи с ними, можно будет делать выводы о качестве работы нейронной сети переводчика и облегчить этим настройку.

Литература

1. Sequence-To-Sequence Models. – URL: https://www.tensorflow.org/tutorials/seq2seq (accessed October 18, 2017).
2. Универсальная грамматика/ Теория Хомского. – URL: https://ru.wikipedia.org/wiki/Универсальная_грамматика (accessed October 18, 2017).
3. Иванова А.Г., Исмагилова Г. К. Новая реальность корпусной лингвистики/ А. Г. Иванова, Г. К. Исмагилова // Информационные технологии в исследовательском пространстве разноструктурных языков: сборник статей I Международной интернет-конференции молодых ученых 5 декабря 2016 года. – Казань, 2017. – С. 21–23.